Patent Abstracts of Japan

PUBLICATION NUMBER : 11328191
PUBLICATION DATE : 30-11-99

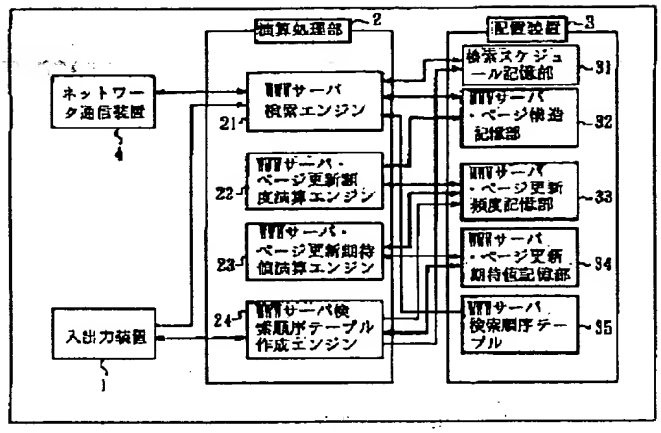
APPLICATION DATE : 13-05-98
APPLICATION NUMBER : 10129829

APPLICANT : NEC CORP;

INVENTOR : KATO TAKASHI;

INT.CL. : G06F 17/30

TITLE : WWW ROBOT RETRIEVING SYSTEM



ABSTRACT : PROBLEM TO BE SOLVED: To provide an automatic leading-out method of a searching reference point not for simple retrieval but for retrieving only a pertinent page by estimating a WWW page updated in advance.

SOLUTION: This system is provided with an update frequency arithmetic engine 22 calculating the update frequency of an optional WWW page from a retrieved result, an update expecting degree arithmetic engine 23 calculating an update expected value in an optional time from the updating frequency and a retrieving order table preparing engine 24 automatically extracting retrieving priority order. A retrieving order table is prepared by estimating whether the optional WWW page is updated at a certain time from the engine 22 and the engine 23. This WWW robot retrieving system retrieves according to this retrieving order table.

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平11－328191

(43)公開日　平成11年(1999)11月30日

(51)Int.Cl.[6]	識別記号	ＦＩ	
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/40	３１０Ｆ
			３８０Ｚ
		15/403	３４０Ｂ

審査請求　有　請求項の数６　ＯＬ　（全７頁）

(21)出願番号　特願平10－129829

(22)出願日　平成10年(1998)５月13日

(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目７番１号

(72)発明者　加藤　剛史
東京都港区芝五丁目７番１号　日本電気株式会社内

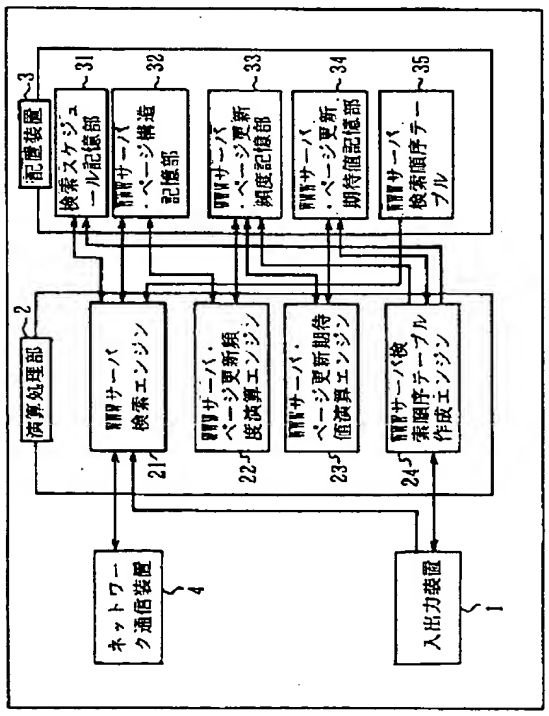
(74)代理人　弁理士　宇高　克己

(54)【発明の名称】　ＷＷＷロボット検索システム

(57)【要約】　（修正有）

【課題】　単純な検索ではなく予め更新されているＷＷＷページを予想して、該当するページだけを検索ための探索基準点の自動導出方法を提供する。

【解決手段】　検索した結果より任意のＷＷＷページの更新頻度を演算する更新頻度演算エンジン２２と、更新頻度から任意の時間における更新期待値を演算する更新期待度演算エンジン２３と、検索優先順位を自動抽出する検索順序テーブル作成エンジン２４を有し、更新頻度演算エンジン２２と更新期待度演算エンジン２３とから、ある時刻において任意のＷＷＷページが更新されているかを予想し、検索順序テーブルを作成する。ＷＷＷロボット検索システムはこの検索順序テーブルに従って検索する。



【特許請求の範囲】
【請求項１】　ＷＷＷページ検索時に自動生成された検索順序に従って検索を実施することを特徴とするＷＷＷロボット検索システム。
【請求項２】　ＷＷＷページ検索時に開始点となり得る複数の探索基準点から最適の基準点を特定し、自動生成された検索順序に従って検索を実施することを特徴とするＷＷＷロボット検索システム。
【請求項３】　ＷＷＷロボットが検索したＷＷＷページの構造を基に任意のＷＷＷページの更新頻度を自動的に抽出することを特徴とするＷＷＷロボット検索システム。
【請求項４】　ＷＷＷページ間のハイパーリンク関係によって各ＷＷＷページから一意に求まる更新頻度から、任意の時間における更新期待度を自動抽出することを特徴とするＷＷＷロボット検索システム。
【請求項５】　各ＷＷＷページの更新頻度と更新期待度からＷＷＷロボットが検索する場合の検索先の優先順序を自動生成することを特徴とするＷＷＷロボット検索システム。
【請求項６】　各ＷＷＷページの更新頻度を演算する更新頻度演算エンジンと、ＷＷＷページの更新頻度から任意の時間における更新期待値を演算する更新期待度演算エンジンと、更新頻度の値と更新期待度の値からＷＷＷロボットが検索する場合の検索優先順位を自動抽出する検索順序テーブル作成エンジンと、を具備することを特徴とするＷＷＷロボット検索システム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、ＷＷＷ（Ｗｏｒｌｄ　Ｗｉｄｅｗｅｂ）ロボットによる検索時の探索基準点の導出方法に関し、特にＷＷＷロボットによる検索を行う場合において、開始点となる複数個の探索基準点から最適な検索順序を求めるためのＷＷＷロボット検出システムに関する。
【０００２】
【従来の技術】ＷＷＷロボット検索システムは、ＷＷＷサーバ内部のＷＷＷページの構造や各ＷＷＷページの更新を検出する機能を有することを特徴とするシステムである。ＷＷＷロボット検索システムは、特定のＷＷＷサーバのトップページやある特定のＷＷＷページを始点としてＷＷＷサーバの検索を実施し、検索によって取得したＷＷＷページ情報からページを記述しているＨＴＭＬ（ＨｙｐｅｒＴｅｘｔＭａｋｅｕｐ　Ｌａｎｇｕａｇｅ）を解析して、このＷＷＷページからハイパーリンクされている次のＷＷＷページの位置を抽出する。なお、ＨＴＭＬは、ＷＷＷにおいてクライアントとサーバとが通信するためのプロトコルであるＨＴＴＰ（ＨｙｐｅｒＴｅｘｔ　Ｔｒａｎｓｆｅｒ　Ｐｒｏｔｃｏｌ）にしたがってハイパーテキストを記述するための言語である。
【０００３】また、この検索システムは、ＷＷＷサーバーを検索していく過程において検出した、新規に作成されたＷＷＷページ、変更されたＷＷＷページ、削除されたＷＷＷページの内容と位置情報を記憶する。ＷＷＷロボットは前述の手順により目的のＷＷＷサーバ内部のＷＷＷページの構造、更新履歴の管理を行う。
【０００４】しかしながら、従来のＷＷＷ検索ロボットにおいては、検索の開始位置は必ず操作者によって予め指定されており、同様に検索順序も操作者があらかじめ指定されている。ＷＷＷ検索ロボットの検索処理は、操作者があらかじめ指定した検索の開始位置と検索順序から、順次ＨＴＭＬを解析しＷＷＷページを取得するルールに基づいて処理している。
【０００５】このように操作者が事前に指定したパラメータをそのままで用いるような検索ルールで動くＷＷＷロボット検索システムにおいては、広範囲に多くのＷＷＷページを取得する場合にＷＷＷロボットが検索する範囲と検索時間、およびネットワークへの負荷は、検索するＷＷＷサーバの数とハイパーリンク（ＨｙｐｅｒＬｉｎｋ）の深さの積算に比例する。
【０００６】従来の文書検索システムにおいて、利用者から指示された語句を含む文書を検索するに当たり、記憶された文書全体を対象として全文検索を行う機能と、各文書から予め抽出された語句により構成される索引を参照して指示語句を含む文書を検索するキーワード検索機能とを備え、さらにこれら両機能のいずれを利用すべきかを、指示された語句その他の条件から判定し、この判定結果にしたがっていずれか有利な検索を行う判定手段を備えた文書検索システムも提案されている（例えば、特開平１０－２１２５５号公報）。しかし、この文書検索システムでは、全文検索、キーワード検索のいずれが有利かについてのみ判定するものであり、検索開始の基準点や検索順序は所定基準に従って実行される。
【０００７】これら従来のＷＷＷロボット検索システムにおいては、必ず操作者の指定した開始点から、順次ハイパーリンクされているＷＷＷページを継続的に検索する。しかし、現在のＷＷＷサーバ内部のＷＷＷページの構造は複雑、かつ、ハイパーリンクの階層も非常に深くなっており、操作者が事前に開始点を与えて、本情報を基に順次検索を行う従来の検索ルールでは、ＷＷＷサーバ内部で実際に変更されたＷＷＷページに到達するまで非常に多くの時間を要するといった問題点が指摘されている。さらに、ＷＷＷサーバ検索はネットワークを経由して行う関係から、不要な検索を多く行うことによりネットワークリソースを浪費し、さらに、ネットワーク負荷を増大させてしまうという問題点もある。
【０００８】
【発明が解決しようとする課題】本発明の課題は、上述のような従来技術の問題点を解消し、ＷＷＷサーバ内の各ＷＷＷページの更新頻度と更新期待度を導出し、これ

ら2つから最適な検索開始点の検索順序を自動的に導出することであり、またWWWロボットによる検索時のネットワークへの負荷を軽減するWWWロボット検索システムを提供することにある。

【0009】

【課題を解決するための手段】本発明は、WWWロボット検索システムがWWWページを検索するための優先順位の決定を自動的に導出して、それに従って実際のWWWページの検索を実施する。より具体的には、図1に示すように、WWWロボット検索システムは、検索した結果より任意のWWWページの更新頻度を演算する更新頻度演算エンジン22と、WWWページの更新頻度から任意の時間における更新期待値を演算する更新期待度演算エンジン23と、更新頻度と更新期待度の値からWWWページの検索優先順位を自動抽出する検索順序テーブル作成エンジン24と、を具備している。

【0010】

【作用】WWWサーバ・ページ構造記憶部32は、WWWサーバ検索エンジン21が検索した結果であるWWWサーバ内部のWWWページの情報とページ間の繋がりを表す構造情報を記憶している。WWWサーバ・ページ更新頻度演算エンジン22は、WWWページの構造情報から各々のWWWページの更新頻度の値を演算し、その結果をWWWサーバ・ページ更新頻度記憶部33に記憶させる。

【0011】WWWサーバ・ページ更新期待度演算エンジン23は更新頻度記憶部33の情報からある時刻における各々のWWWページが更新される期待度を演算するためのパラメータを自動生成して、このパラメータをWWWサーバ・ページ更新期待度記憶部34に記憶させる。WWWサーバ検索順序テーブル作成エンジン24は更新頻度と更新期待度から次回、WWWサーバ検索エンジン21が検索を実行する際に、どのWWWサーバのページから検索しているかを順序付けたテーブルを自動生成する。

【0012】

【発明の実施の形態】次に、本発明の実施の形態について図面を参照して詳細に説明する。図1を参照すると、本発明の第一の実施の形態は、キーボートやディスプレイなどの入出力装置1とプログラム制御により動作する演算処理部2と、情報を記憶する記憶装置3、そしてインターネット等を介して外部のWWWサーバと情報のやり取りを行うネットワーク通信装置4とを含む。

【0013】記憶装置3は、検索スケジュール記憶部31と、WWWサーバ・ページ構造記憶部32、WWWサーバ・ページ更新頻度記憶部33、WWWサーバ・ページ更新期待値記憶部34、検索順序テーブル35とを備える。

【0014】検索スケジュール記憶部31は、過去にWWWサーバを検索した開始時間と検索に要した時間の履歴情報を記憶する。

【0015】WWWサーバ・ページ構造記憶部32は、検索することによって得られたWWWサーバ内に配置されているWWWページの文章の内容とこれらWWWページが各々どのような接続関係にあるのかを記憶する。

【0016】WWWサーバ・ページ更新頻度記憶部33は、各WWWページの更新頻度の度合いを算出して数値化した情報を記憶する。

【0017】WWWサーバ・ページ更新期待値記憶部34は、各WWWページが任意の時間の時点で更新されていると期待される可能性を算出して数値化した情報を記憶する。

【0018】検索順序テーブル35は、次回に検索を行う場合に、どのWWWサーバのどのWWWページから検索するのかという観点から優先順序づけられた情報を記憶する。

【0019】演算処理部2は、WWWサーバ検索エンジン21と、WWWサーバ・ページ更新頻度演算エンジン22、WWWサーバ・ページ更新期待値演算エンジン23、WWWサーバ検索順序テーブル作成エンジン24とを備える。

【0020】WWWサーバ検索エンジン21は、入出力装置1からの実行命令を契機として検索順序テーブルが記憶している順序情報に従って、ネットワーク装置4を経由して外部のWWWサーバ・ページの検索を実施する。検索の結果、WWWページの更新や新規WWWページの追加・削除、およびWWWページ間のハイパーリンク関係といった情報をWWWサーバ・ページ構造記憶部32に記憶させる。WWWページを検索した時間の情報は検索スケジュール記憶部31に記憶させる。

【0021】WWWサーバ・ページ更新頻度演算エンジン22は、WWWサーバ・ページ構造記憶部32が記憶しているWWWページの更新情報やハイパーリンク情報とWWWサーバ・ページ更新頻度記憶部33が記憶しているWWWページ毎の更新頻度に関する情報を基に新たに各々のWWWページの更新頻度を演算して、演算結果をWWWサーバ・ページ更新頻度記憶部33に記憶させる。

【0022】WWWサーバ・ページ更新期待値演算エンジン23は、WWWサーバ・ページ更新頻度記憶部33が記憶している情報を元に、各WWWページが任意の時間の時点で更新されていると期待される度合いを演算して、その結果をWWWサーバページ更新期待値記憶部34に記憶させる。

【0023】WWWサーバ検索順序テーブル作成エンジン24は、入出力装置1から入力されたキーとWWWサーバ・ページ更新頻度記憶部33、およびWWWサーバ・ページ更新期待度記憶部34が記憶している情報から検索順序を演算して、その結果を検索順序テーブル35に記憶させる。

【００２４】
【動作の説明】次に、図１、図２、図３、図４、図５を参照して本発明にかかる構成の動作について詳細に説明する。
【００２５】入出力装置１からの実行指示を受信を契機に、WWWサーバ検索エンジン２１は検索順序テーブル３５があらかじめ記憶している検索順序に従ってネットワーク通信装置４を用いて外部のWWWサーバからWWWページの情報を取得する。検索によって取得するWWWページ情報は、「現在検索しているWWWページを記述しているＨＴＭＬ」、「現在検索しているWWWページが前回検索したときから更新されているか」、「現在検索しているWWWページをハイパーリンクしていた親のWWWページ」、「現在検索しているWWWページがハイパーリンクしている子のWWWページ」の４つである。WWWサーバ検索エンジン２１はこれら４つのWWWページに関する情報をWWWサーバ・ページ構造記憶部３２に記憶させる。
【００２６】WWWサーバ・ページ更新頻度演算エンジン２２はWWWサーバ・ページ構造記憶部３２が記憶している情報を基に、WWWサーバやWWWページの更新の頻度の度合いを示すWWWサーバ・ページ更新頻度を演算して決定する。
【００２７】WWWサーバ検索エンジン２１で検索したWWWサーバ内部のWWWページの構造は図２のようなパーセプトロン型のニューラルネットと同等の形状をしている。そこで、本発明では各WWWページのニューラルネットのノードと見なし、更新頻度を各WWWページ（ノード）の持つ重みがWWWページの更新頻度に相当すると考えて各WWWページの更新頻度を演算していく。図２のようなWWWページ構造をしているWWWサーバの場合、WWWサーバ・ページ更新頻度演算エンジン２２は任意のWWWページa の更新頻度をWWWページa からハイパーリンクする子のWWWページの更新頻度と現状のWWWページa の更新頻度の値から図２に示す式(1) を用いて求める。
【００２８】WWWサーバ・ページ更新頻度演算エンジン２２はWWWサーバ内の全てのWWWページの更新頻度を導出する手法として、任意のWWWページに着目して、その着目したWWWページの更新頻度を演算していく手法を用いる。着目するWWWページの決定手順を以下で説明していく。
【００２９】第一に、WWWサーバ・ページ更新頻度演算エンジン２２は、WWWサーバ・ページ構造記憶部３２が記憶しているWWWページの中で最下部に位置する子のWWWページに着目して、このページのWWWサーバ・ページ更新頻度を導出する。第二に、先ほど更新頻度を導出したページにマークを付ける。第三に、WWWサーバ・ページ更新頻度演算エンジン２２は、WWWサーバ・ページ構造記憶部３２が記憶しているWWWページの中でマークの付いていないWWWページの中で最下部に位置する子のWWWページを次の着目するWWWページとする。以後、この手順を繰り返すことによりWWWページの更新頻度を導出していく。
【００３０】８つのWWWページが図３のような関係でハイパーリンクされている場合には、演算順序は、例えば図３の枠１内に示すように『ページ８→ページ５→ページ６→ページ７→ページ２→ページ３→ページ４→ページ１』という具合に演算していく。
【００３１】WWWサーバ・ページ更新頻度演算エンジン２２は上記で述べた手法により各々のWWWページの更新頻度を導き出し、求めた更新頻度をWWWサーバ・ページ更新頻度記憶部３３に記憶させる。
【００３２】WWWサーバ・ページ更新期待度演算エンジン２３は、ある時点におけるWWWページの更新されていると期待できる度合いの値を演算して、その結果をWWWサーバ・ページ更新期待度記憶部３４に記憶させる。
【００３３】WWWサーバ・ページ更新期待度エンジン２３は、各々のWWWページの更新期待度を求める手法として次の手法を用いて演算する。WWWサーバ・ページ更新期待度演算エンジン２３はWWWページの更新期待を演算する場合に、検索スケジュール記憶部３１が記憶している前回検索を実施した時点から現在検索を実施した時点に至るまでの経過時間と、WWWサーバ・ページ更新頻度記憶部３３が記憶している更新頻度wを用いて、更新期待度Ex を導き出すための図４の式（１）のパラメータεを図４の式（２）を用いて演算する。任意の時刻tにおける更新期待度Ex は図４の式（１）から導出する。
【００３４】WWWサーバ検索順序テーブル作成エンジン２４がWWW検索ロボットの検索順序情報を作成する手順は以下のとおりである。WWWサーバ検索順序テーブル作成エンジン２４は、第一にWWWサーバ・ページ構造記憶部３２が記憶しているWWWページ情報を検索順序テーブル３５へ複写する（図５の表5-1 ）。
【００３５】第二に、WWWサーバ検索順序テーブル作成エンジン２４は、WWWサーバ・ページ更新期待度記憶部３４、および検索スケジュール記憶部３１が記憶している情報から、各WWWページの更新確信度を演算して、その結果を検索順序テーブル３５に記憶させる（図５の表5-2 ）。
【００３６】第三にWWWサーバ検索順序テーブル作成エンジン２４は、演算によって求めた各WWWページの更新期待度を値の大きいものが検索の優先順位の高いものと見なして、検索順序テーブル３５の情報を優先順序の高い順に並び替える（図５の表5-3 ）。
【００３７】第四にWWWサーバ検索順序テーブル作成エンジン２４は、入出力装置１を介して人間（操作者）から与えられた条件に従って、検索順序テーブル３５が

記憶している情報の中から条件に合わないＷＷＷページ情報を削除する（図５の表5-4 ）。

【００３８】

【発明の効果】本発明にかかるＷＷＷロボット検索システムによって得られる効果は、対象としているＷＷＷページとそのＷＷＷページがハイパーリンクする子のＷＷＷページの更新状態から、対象としているＷＷＷページのある時刻における更新期待度を自動生成して、その結果を基にして次回の検索時の優先順序を決定することである。これにより、ＷＷＷロボットは更新されている期待度の高いＷＷＷページから先に検索していくことになるので、更新されているＷＷＷページの情報をすばやく取得することが可能となる。

【００３９】また、検索テーブルを作成時に人間（操作者）の与える条件に従って検索するに値しないＷＷＷページを間引きできる（図５の表5-4 ）ので、不要な検索作業を軽減させて、ＷＷＷロボット検索システムが検索する際に発生するネットワークへの負荷を軽減することができる。

【図面の簡単な説明】

【図１】本発明の実施の形態の構成を示すブロック図である。

【図２】本発明の実施の形態の動作の更新頻度の演算方法を示す説明図である。

【図３】本発明の実施の形態の動作の更新頻度演算時の演算順序の具体例を示す説明図である。

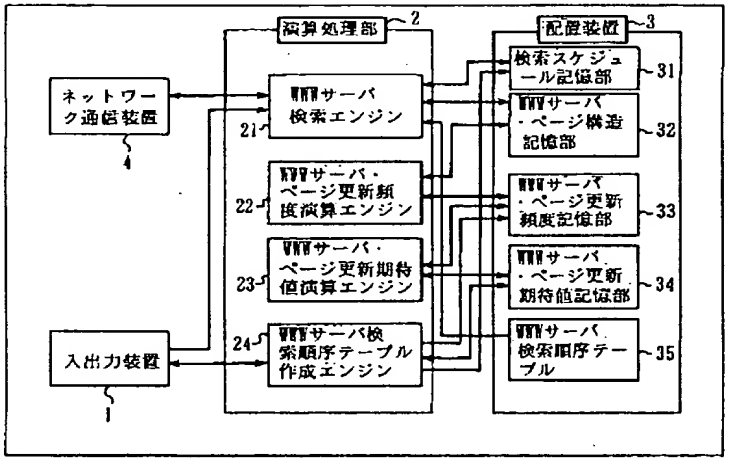
【図４】本発明の実施の形態の動作の更新期待度の演算方法を示す説明図である。

【図５】本発明の実施の形態の動作の検索テーブル作成方法を示す説明図である。

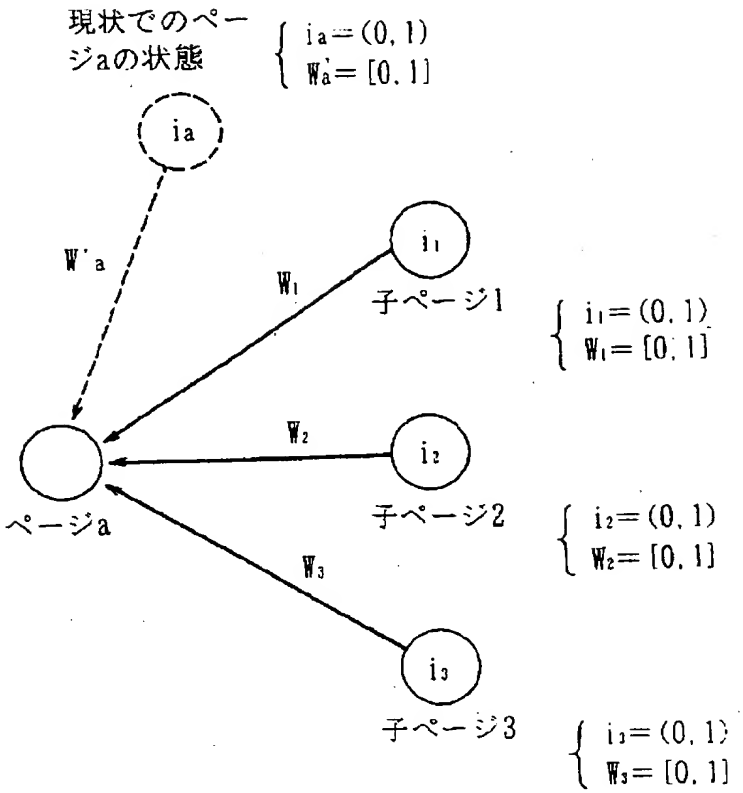
【符号の説明】

１　入出力装置
２　演算処理部
３　記憶装置
４　ネットワーク通信装置
２１　ＷＷＷ　サーバ検索エンジン
２２　ＷＷＷ　サーバ・ページ更新頻度演算エンジン
２３　ＷＷＷ　サーバ・ページ更新期待度演算エンジン
２４　ＷＷＷ　サーバ検索順序テーブル作成エンジン
３１　検索スケジュール記憶部
３２　ＷＷＷ　サーバ・ページ構造記憶部
３３　ＷＷＷ　サーバ・ページ更新頻度記憶部
３４　ＷＷＷ　サーバ・ページ更新期待度記憶部
３５　ＷＷＷ　サーバ検索順序テーブル

【図１】

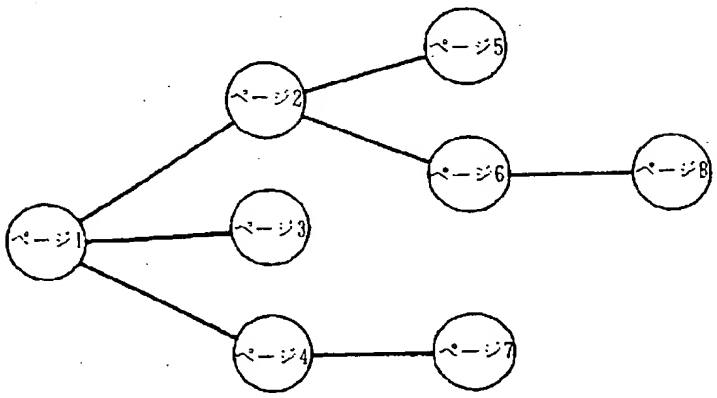


【図2】



$$w_a = \beta(1-i_a)w'_a + f(w'_a i_a + \sum^{n} w_n i_n - \theta) \quad \cdots\cdots(1)$$

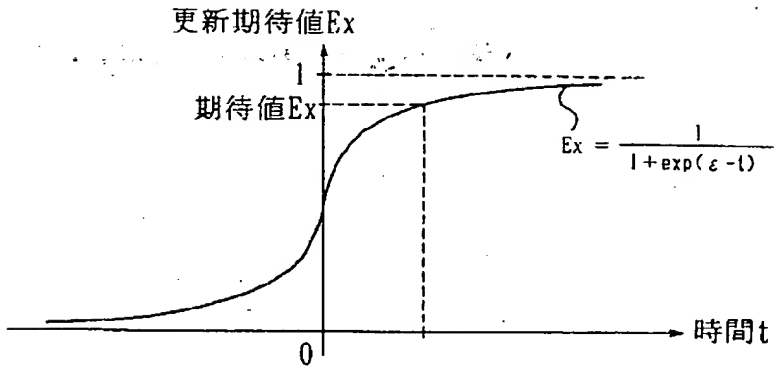
【図3】



演算していく順序は以下のとおり。

ページ8→ページ5→ページ6→ページ7
→ページ2→ページ3→ページ4→ページ1 ……1

【図4】



時刻 t_a においてWWWページ a が更新される期待値 Ex_a は以下の式から求める。

$$Ex_a = \frac{1}{1+\exp(\varepsilon_a - t_a)} \quad \cdots\cdots (1)$$

ε_a はWWWページ a の更新頻度 w_a、WWWサーバ検索の検索時間間隔の平均値 T_{avg} から以下の式を用いて求める。

$$\varepsilon_a = T_{avg} + \log_2 (\frac{1}{w_a} - 1) \quad \cdots\cdots (2)$$

【図5】